# Creating Public Health Standard Vocabularies: Mapping A Set of CDC's Pathogen Codes to SNOMED Concepts

**Wenkai Li, MD, MAS, MS[1,2], James Tolson, BS[1], Teresa C. Horan, MPH[1]**
**[1] Division of Healthcare Quality Promotion, NCID, CDC, Atlanta, GA, USA**
**[2] Public Health Informatics Fellowship Program, DPHSI, EPO, CDC, Atlanta, GA, USA**

**Abstract**
CDC and its public health partners are conducting pilot projects designed to capture data directly from healthcare providers or to exchange existing data electronically.[1] The translation of proprietary standards into the accepted national standards is a key part of this effort. This study mapped a set of pathogens within CDC's hospital infection control activities to The Systematized Nomenclature of Medicine (SNOMED) concepts and investigated the differences between the two.

**Background**
Since the 1950s, CDC has been investigating nosocomial infections and outbreaks in healthcare facilities. To promote healthcare and patient safety, CDC is taking the lead in developing the National Healthcare Safety Network (NHSN), an internet-based reporting system. The NHSN will unite three existing surveillance systems: the National Nosocomial Infection Surveillance (NNIS) system, the National Surveillance System for Healthcare Workers (NaSH), and the Dialysis Surveillance Network (DSN).[2] NHSN will implement electronic messaging as the new communication method between CDC and participating institutions. Achieving this goal requires the use of standards for messaging health information. SNOMED is one of the most widely used vocabularies in healthcare. We created and implemented methods to map a set of CDC's NNIS pathogen codes (a proprietary standard) to SNOMED concepts (a standard vocabulary).

**Methods**
The mapping process consists of three sequential steps. First, we created SQL programs to automatically match the names of NNIS pathogens against those of SNOMED organisms. Second we used a freeware browser and terminology server that supports SNOMED Clinical Terms (CT) to check the results from the automatic matching and then manually searched unmatched items against those in the SNOMED database. Finally, we consulted with epidemiologists and lab scientists on the remaining unmatched items.
To perform the automatic matching by pathogen names, we attempted to match the first two words of pathogen names from the two standards. We used the SQL server 2000, SNOMED CT database and NNIS pathogen list in an ASCII file to implement the mapping. The critical step in the automatic mapping process is subsetting the list of infectious agents from SNOMED database. We used the IS-A hierarchy in the SNOMED CT database and applied the recursive algorithm to construct a list of children of a single concept with SNOMED Concept ID 36272005 (infectious agent).
To preserve the integrity of the NNIS pathogen list in the matching result, we used SQL queries to left-join the NNIS pathogen list with SNOMED infectious agent list.

**Results**
There were 12,960 unique concepts in our subset of SNOMED infectious agents. The NNIS standard has 1,214 pathogen codes. Of these, 170 were eliminated because they were ambiguous or duplicates. Of the 1,044 remaining NNIS Codes, 987 (94.5%) matched a SNOMED concept and 57 could not be matched with SNOMED concepts. We are consulting with lab scientists to provide solutions to the remaining mismatches. Of the 12,960 SNOMED infectious agents, 11,973 (92.4%) did not match with a NNIS code. Some of them will be evaluated to determine whether they should be added to the NNIS pathogen list.

**Conclusions**
This study provided us a good understanding of both the proprietary standard (NNIS pathogen codes) and the standard vocabulary (SNOMED CT). It is also a first step towards creating public health standard vocabularies. We will continue to work with standards development organizations to propose and enhance the standard vocabularies to accommodate public health needs.

**References**
1. Koo D, Meade M, Broome CV. New Means of Data Collection. In: O'Carroll PW, et al, eds: Public Health Informatics and Information Systems. New York, NY: Springer-Verlag, New York, Inc.; 2003.
2. CDC. The National Healthcare Safety Network (NHSN) Protocol, Version 11. October 2002; Atlanta, GA (unpublished).